دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد

عنوان:

مقایسه روش ماشین بردار پشتیبان با روش‌های رگرسیون لجستیک و تحلیل ممیزی

در طبقه‌بندی روابط جنسی خارج از ازدواج دائم در جوانان ۲۹-۱۹ ساله ایران

توسط: شهره شفیعی

استاد راهنما:

دکتر محمدرضا بانشی

اساتید مشاور: دکتر حمید شریفی

میترا منتظری

سال تحصیلی ۹۷-۱۳۹۶

Kerman University of Medical Sciences

School of Health

In Prtial Fulfillment of the Requirements for the Degree Master of Science

Title:

Comparison of Support Vector Machine, Logistic Regression and Discriminant Analysis in Classification of Having Extramarital Sexual Contacts in Iranian Youth ۱۹-۲۹ Years Old

By:

Shohreh Shafiei

Supervisors:

Dr. Mohammad Reza Baneshi

Adviser:

Dr. Hamid Sharifi

Mitra Montazeri

Year: ۲۰۱۸

مقدمه: یکی از مهم‌ترین کاربردهای روش‌های آماری در علوم مختلف طبقه‌بندی است. رگرسیون لجستیک و تحلیل ممیزی از پرکاربردترین روش‌ها در مسائل طبقه‌بندی هستند. استفاده از ماشین بردار پشتیبان، رویکرد جدیدی است که در چند سال اخیر مورد توجه بسیاری قرار گرفته است. این روش از جمله روش‌هایی است که پیش‌فرض قابل توجهی ندارد. هدف از این مطالعه مقایسه رگرسیون لجستیک، تحلیل ممیزی و ماشین بردار پشتیبان برای طبقه‌بندی روابط جنسی خارج از ازدواج دائم در جوانان ۲۹-۱۹ ساله ایران می‌باشد.

روش‌ها: داده‌ها شامل ۲۹۹۲ نفر جوان ۱۹ تا ۲۹ ساله بوده و تعداد متغیرهای مستقل ۱۴ بود. متغیر روابط جنسی خارج از ازدواج دائم به عنوان متغیر پاسخ در نظر گرفته شد. بعد از بررسی داده‌ها و برآورد مقادیر گمشده روش‌های رگرسیون لجستیک و تحلیل ممیزی و ماشین بردار پشتیبان به داده‌ها برازش داده شد و پیش‌بینی روابط جنسی خارج از ازدواج دائم براساس این روش‌ها انجام شد. برای تعیین تاثیر حجم نمونه بر روی روش‌ها با کاهش حجم نمونه به ۱۵۰۱، ۷۴۸ و ۴۰۲ نفر، روش‌ها مقایسه شدند. جهت محاسبه میزان حساسیت، ویژگی، دقت و سطح زیر منحنی راک برای مقایسه قدرت پیش-بینی مدل ها از نرم افزار R استفاده گردید.

یافته‌ها: در حالتی که حجم نمونه ۲۹۹۲ نفربود، حساسیت برای روش‌های رگرسیون لجستیک، تحلیل ممیزی و ماشین بردار پشتیبان با کرنل تابع پایه شعاعی و چند جمله‌ای به ترتیب ۰/۲۳، ۰/۳۵، ۰/۹۹ و ۰/۱۵ بود. ویژگی به ترتیب ۰/۹۶، ۰/۸۹، ۱ و ۰/۹۹ بود. دقت نیز برابر با ۰/۸۱، ۰/۷۸، ۰/۹۸ و ۰/۸۲ بود. همچنین سطح زیر منحنی راک به ترتیب برابر با ۰/۶۰، ۰/۷۵، ۰/۹۹ و ۰/۷۷ بود. در حالت با حجم ۱۵۰۱ نفر، حساسیت برای رگرسیون لجستیک، تحلیل ممیزی، ماشین بردار پشتیبان با کرنل تابع پایه شعاعی و چند جمله‌ای برابر با ۰/۲۴، ۰/۳۷، ۰/۹۷، ۰/۲۲ و ویژگی برابر با ۰/۹۷، ۰/۸۹، ۱،

۱ و دقت برابر با ۰/۸۲، ۰/۷۹، ۰/۹۹، ۰/۸۴ و سطح زیر منحنی راک به ترتیب برابر با۰/۶۰، ۰/۷۵، ۱، ۰/۸۳ بود در حالت با حجم نمونه ۷۴۸ نفر، حساسیت برای رگرسیون لجستیک، تحلیل ممیزی، ماشین بردار پشتیبان با کرنل تابع پایه شعاعی و چند جمله‌ای برابر با ۰/۳۱، ۰/۴۲، ۰/۹۳، ۰/۲۳ و ویژگی برابر با ۰/۹۶، ۰/۸۹، ۱، ۱ و دقت برابر با ۰/۸۳، ۰/۸۰، ۰/۹۸، ۰/۸۱ بود. و سطح زیر منحنی راک نیز به ترتیب برابر با ۰/۶۴، ۰/۸۰، ۱، ۰/۸۷ بود. در حالت با حجم نمونه ۴۰۲ نفر، حساسیت برای رگرسیون لجستیک، تحلیل ممیزی، ماشین بردار پشتیبان با کرنل تابع پایه شعاعی و چند جمله‌ای برابر با ۰/۲۹، ۰/۴۵، ۰/۸۲، ۰/۰۷ و ویژگی برابر با ۰/۹۶، ۰/۹۲، ۱، ۱ و دقت برابر با ۰/۸۲، ۰/۸۳، ۰/۹۶، ۰/۸۱ بود. سطح زیر منحنی راک به ترتیب برابر با ۰/۶۳، ۰/۸۴، ۱، ۰/۸۶ بود.

نتیجه‌گیری: یافته ها نشان دادند که مقادیر حساسیت، ویژگی، دقت و سطح زیر منحنی راک برای ماشین بردار پشتیبان با کرنل تابع پایه شعاعی در حجم‌های متفاوت نمونه، بیشتر از تابع کرنل چندجمله‌ای و دو روش دیگر بود.

کلمات کلیدی: رگرسیون لجستیک، تحلیل ممیزی، ماشین بردار پشتیبان، حساسیت، ویژگی، سطح زیر منحنی راک

Abstract

Introduction: Classification is one of the most important applications of statistical methods. Logistic Regression (LR) and Discriminant Analysis (DA) are applied in most settings. Using of Support Vector Machine (SVM) as a modern modeling method has received considerable attention in recent years. This method does not depend on pre-assumption.

The aim of this study is to compare the LR, DA and SVM for classification of Iranian youth ١٩-٢٩ years old experiencing extramarital sexual contacts.

Methods: Information of ١٤ independent variables for ٢٩٩٢ individuals aged ١٩ to ٢٩ years were collected. Extramarital sexual contacts considered as the response variable. After imputation of missing values, LR, DA, SVM was fitted. To determine the effect of sample size methods were compared with reduce sample size to ١٥٠١, ٧٤٨, ٤٠٢. Finally measures of the sensitivity, specificity, accuracy and area under ROC curve was applied to compare the methods using R software.

Result: At the sample size of ٢٩٩٢, the sensitivity in LR, DA, SVM with radial basis function and polynomial was equal to ٠,٢٣, ٠,٣٥, ٠,٩٩, ٠,١٥ respectively. The specificity was equal ٠,٩٦, ٠,٨٩, ١ and ٠,٩٩ respectively. The accuracy was ٠,٨١, ٠,٧٨, ٠,٩٨, ٠,٨٢. The area under ROC was ٠,٦٠, ٠,٧٥, ٠,٩٩, ٠,٧٧. N=١٥٠١ sensitivity in LR, DA, SVM with kernel function radial basis function and polynomial was equal ٠,٢٤, ٠,٣٧, ٠,٩٧, ٠,٢٢ respectively. Specificity was equal ٠,٩٧, ٠,٨٩, ١, ١ respectively. Accuracy was ٠,٨٢, ٠,٧٩, ٠,٩٩, ٠,٨٤ . Area uder ROC was ٠,٦٠, ٠,٧٥, ١, ٠,٨٣. When N=٧٤٨ sensitivity in LR, DA, SVM with radial basis function and polynomial was equal ٠,٣١, ٠,٤٢, ٠,٩٣, ٠,٢٣ respectively. Specificity was equal ٠,٩٦, ٠,٨٩, ١, ١ respectively. Accuracy was ٠,٨٣, ٠,٨٠, ٠,٩٨, ٠,٨١. Area uder ROC was ٠,٦٤, ٠,٨٠, ١, ٠,٨٧. When

N=٤٠٢ sensitivity in LR, DA, SVM with radial basis function and polynomial was equal ٠,٢٩, ٠,٤٥, ٠,٨٢, ٠,٠٧ respectively. Specificity was equal ٠,٩٦, ٠,٩٢, ١, ١ respectively. Accuracy was ٠,٨٢, ٠,٨٣, ٠,٩٦, ٠,٨١. Area uder ROC was ٠,٦٣, ٠,٨٤, ١, ٠,٨٦.

Conclusion: The result suggested the superiority of SVM with radial basis function.

Keywords: logistic regression, discriminant analysis, support vector machine, sensitivity, specificity, ROC